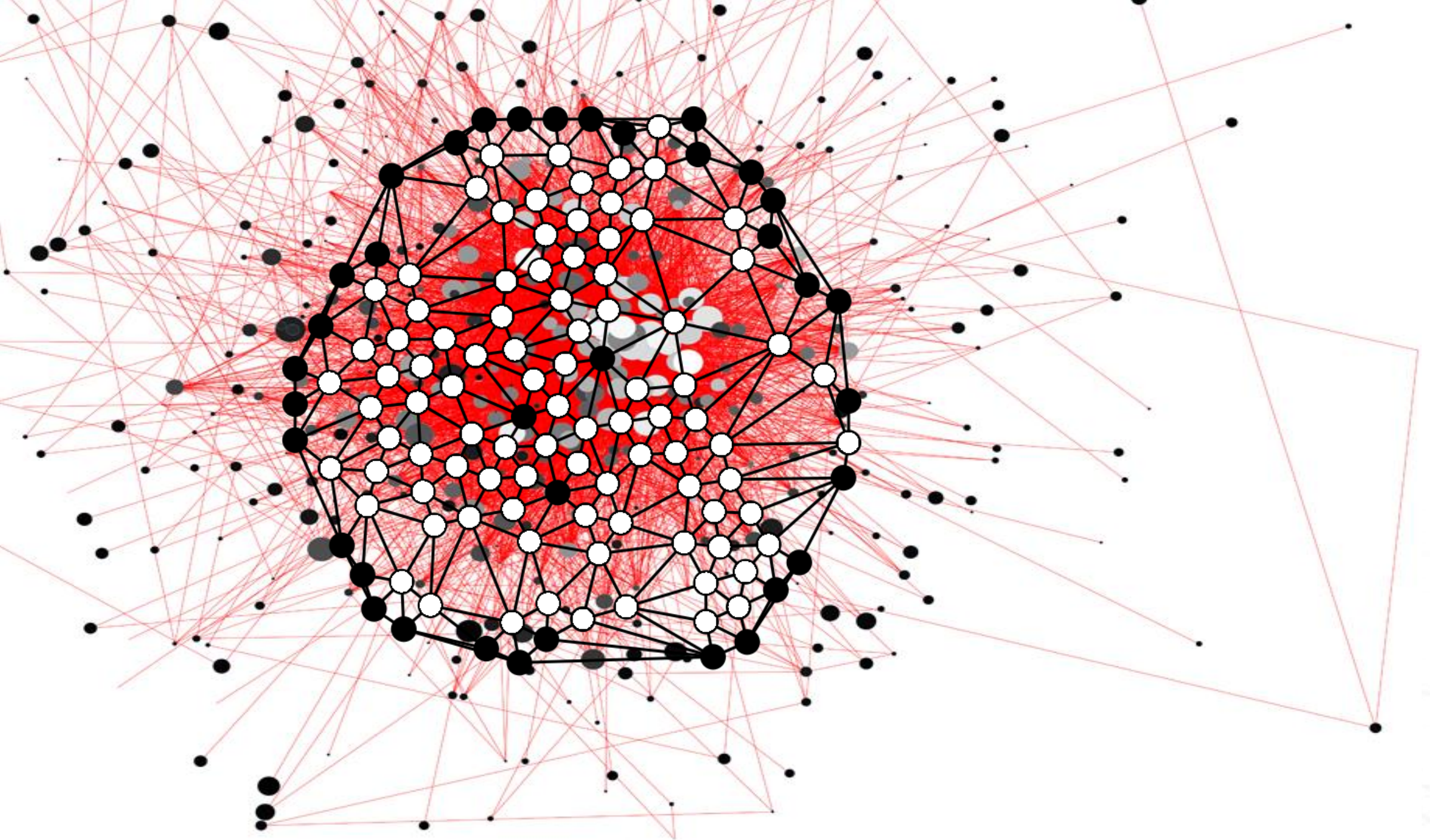
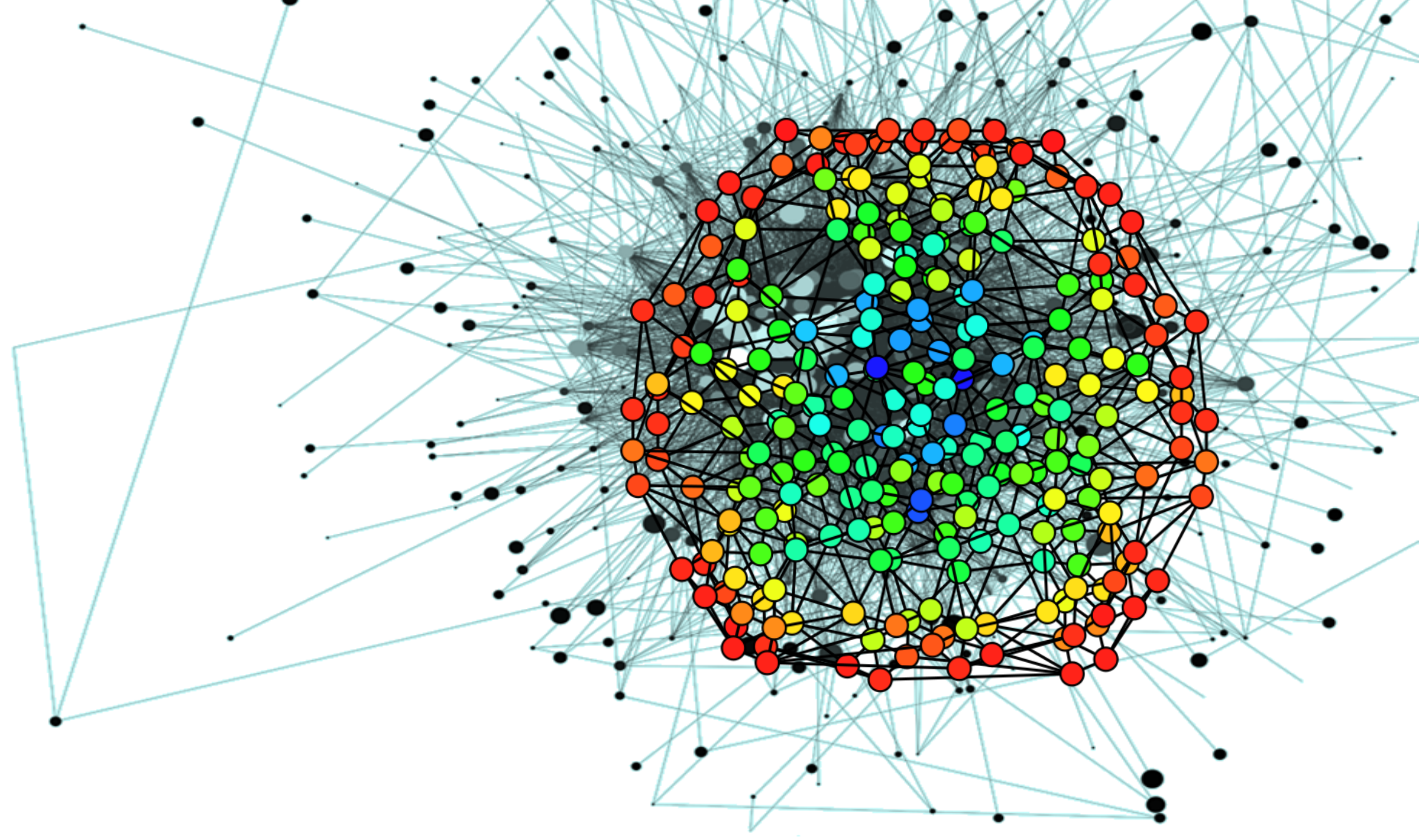
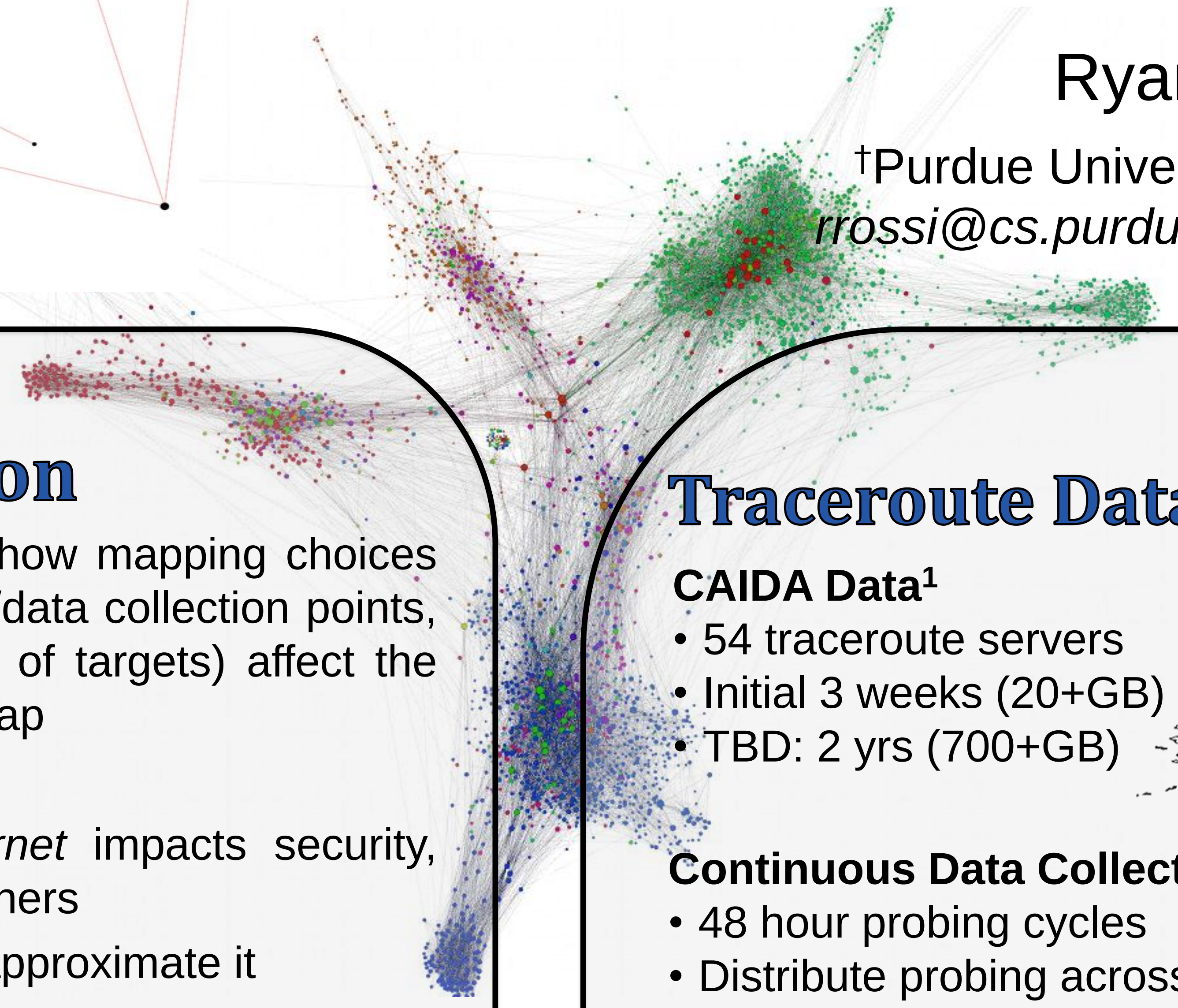




# Approximating Coverage of Internet Maps
## From Multiple Vantage Points

Ryan Rossi† and Brian Gallagher*

†Purdue University    *Lawrence Livermore National Laboratory
rrossi@cs.purdue.edu    bgallagher@llnl.gov

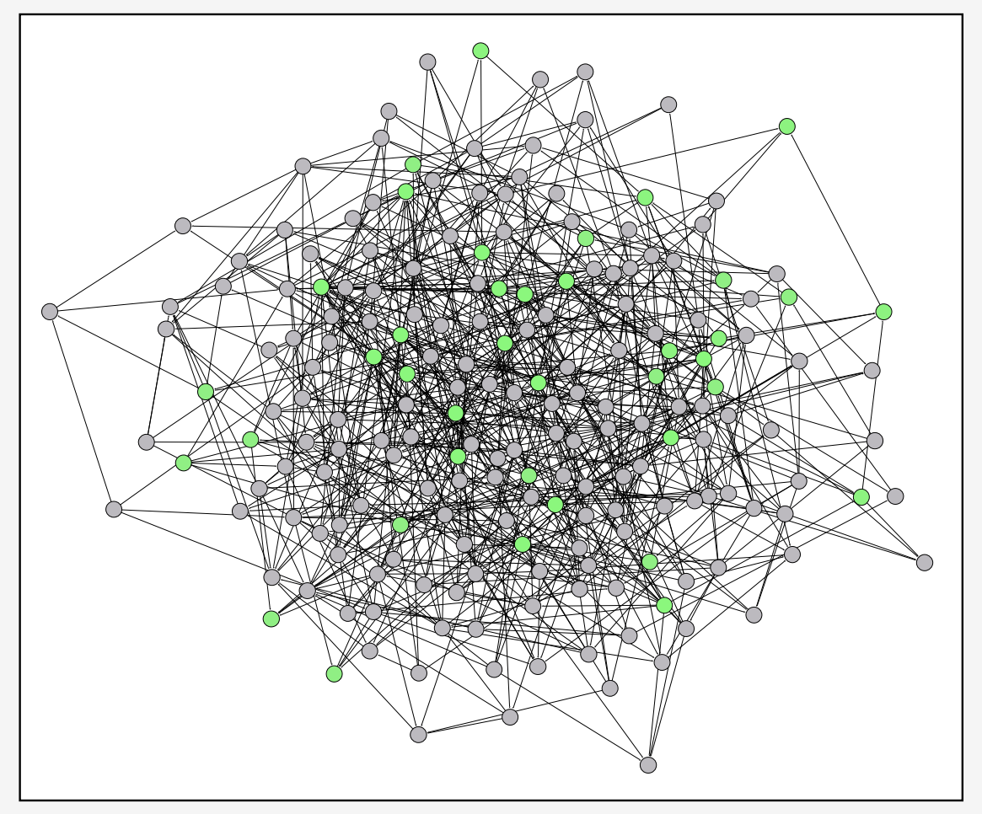


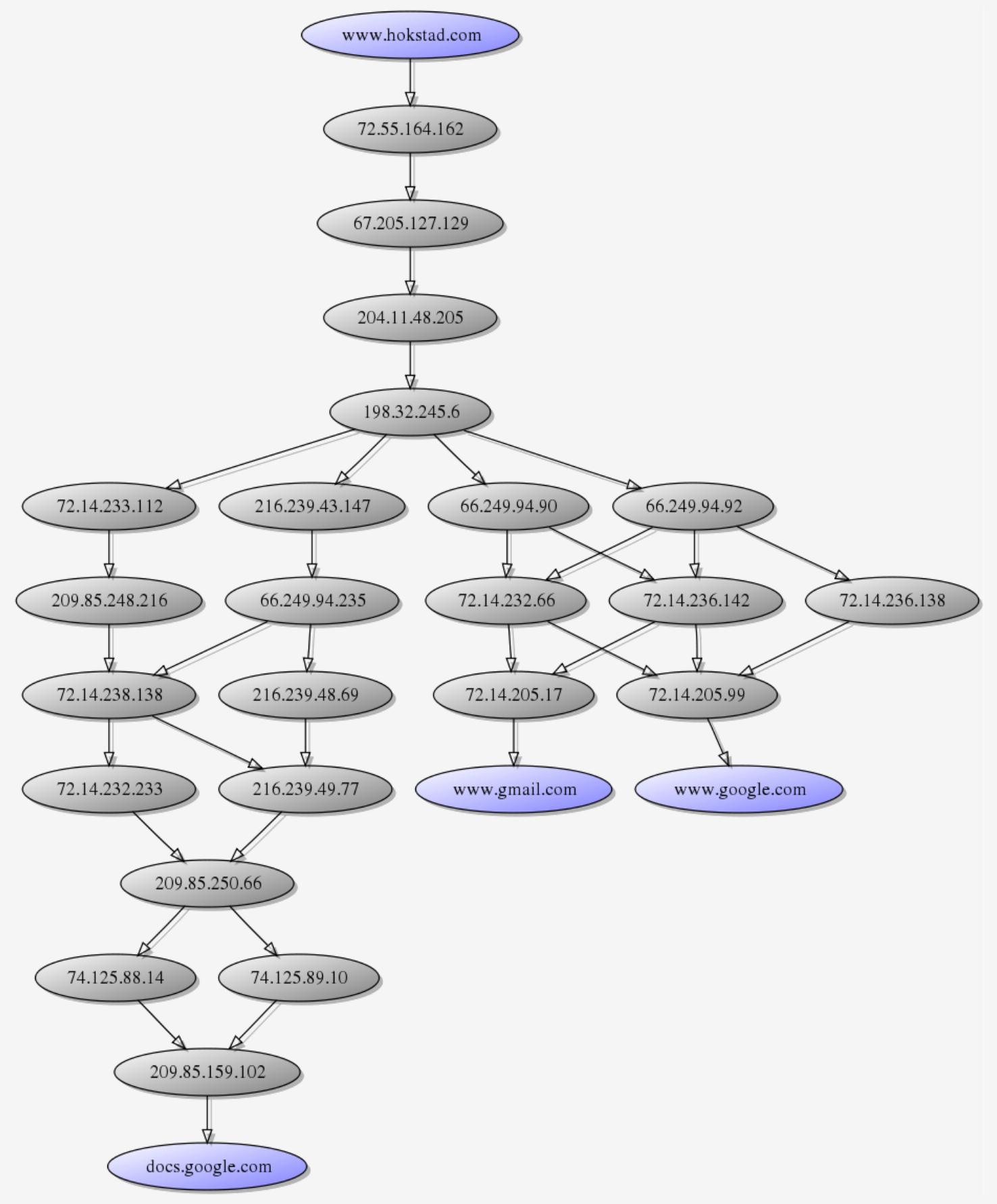
## Problem & Motivation

**Problem:** We want to understand how mapping choices (e.g., number of traceroute servers/data collection points, amount of data over time, number of targets) affect the coverage of the resulting network map

**Motivation:** *Structure of the Internet* impacts security, performance, robustness, among others

- Impossible to observe, can only approximate it
- Conclusions are made from these approximations
- Important to understand the quality of these approximations and the factors that influence them





## Challenges

BIG Data

Data collection problems:

•Observability issues
•Asymmetric routing
•Topology changes
•BGP costs/incentives evolve
•“Hot potatoe” routing
•Multiple IPs for router (ambiguous)
•…

**Impossible** to observe the actual internet map!
- must approximate!

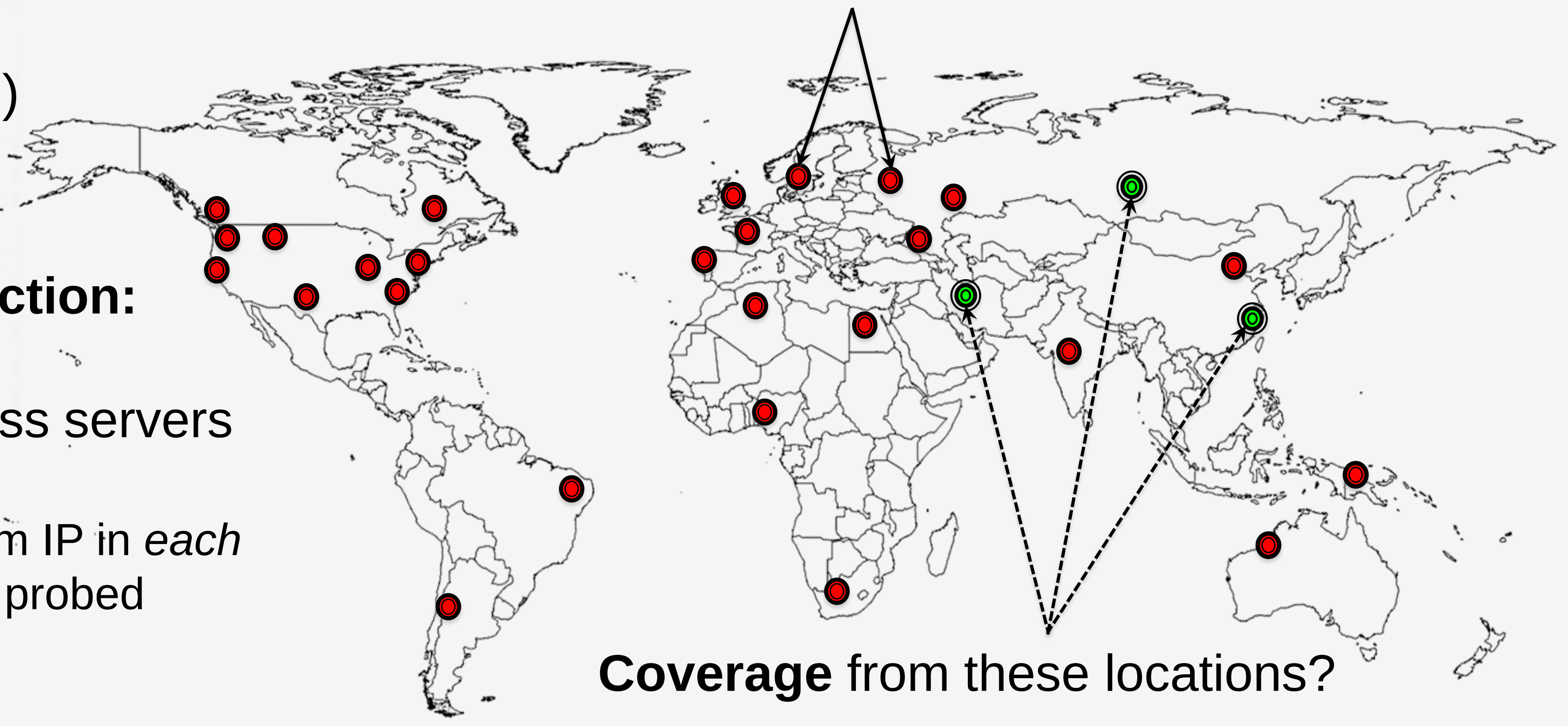
## Traceroute Data & Collection

**CAIDA Data**[1]
• 54 traceroute servers
• Initial 3 weeks (20+GB)
• TBD: 2 yrs (700+GB)

**Continuous Data Collection:**
• 48 hour probing cycles
• Distribute probing across servers

000.000.000.*
⋮
255.255.255.*
] A random IP in *each prefix* is probed

**Data collection points** (traceroute servers)

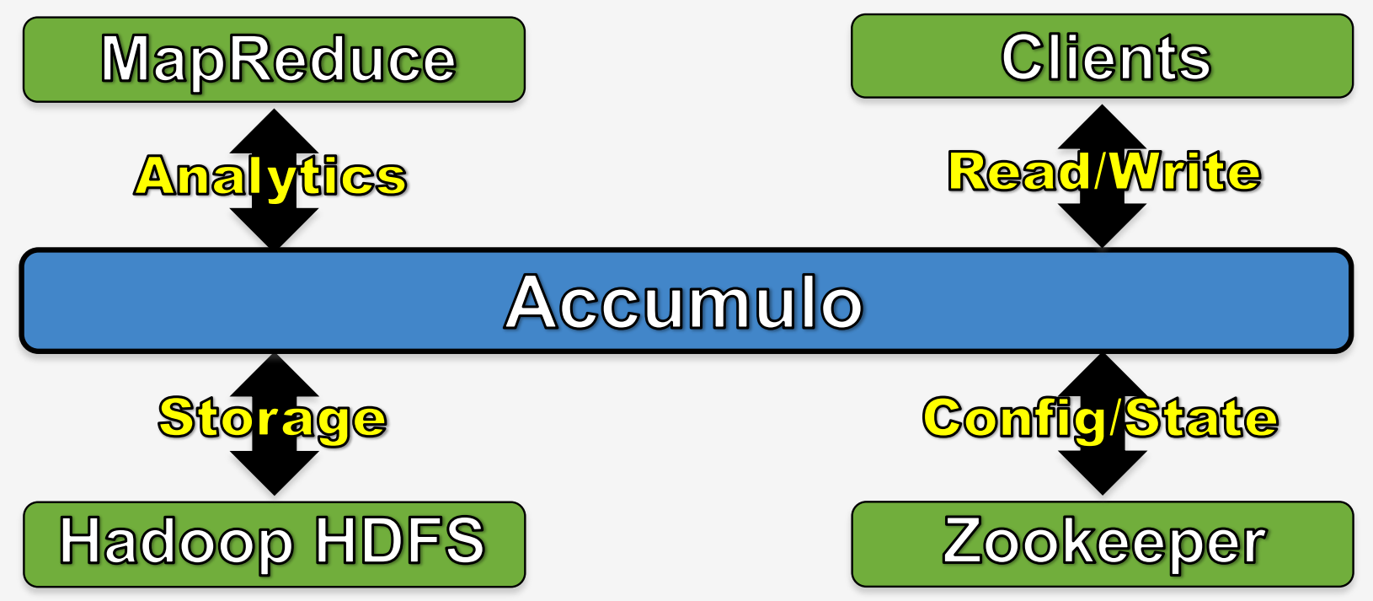


**Coverage** from these locations?

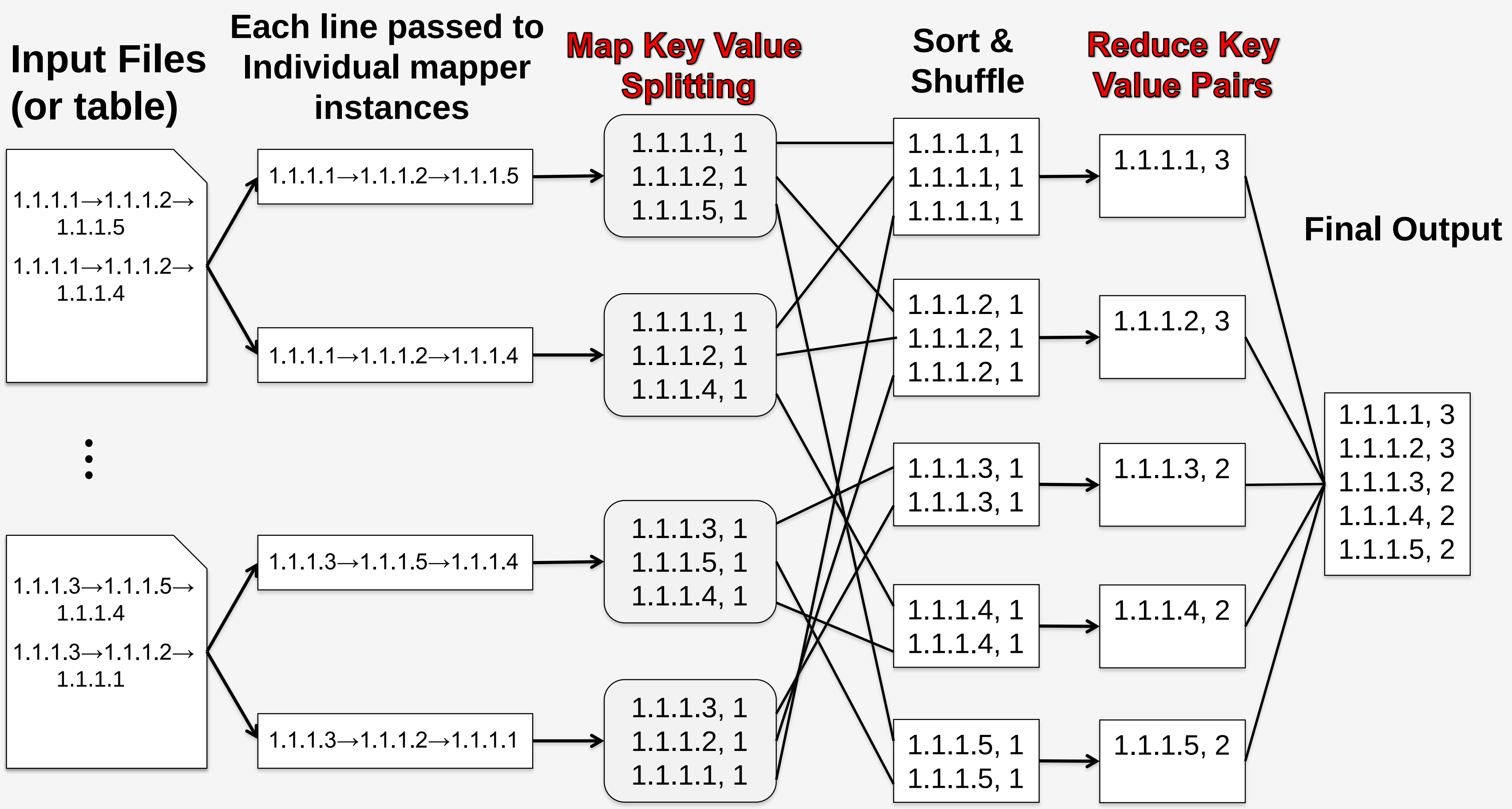
## Framework for Analyzing Real-time Internet Coverage

**Accumulo:** Scalable- distributed key-value store, enables interactive access to trillions of records, petabytes of data across 1000’s of servers

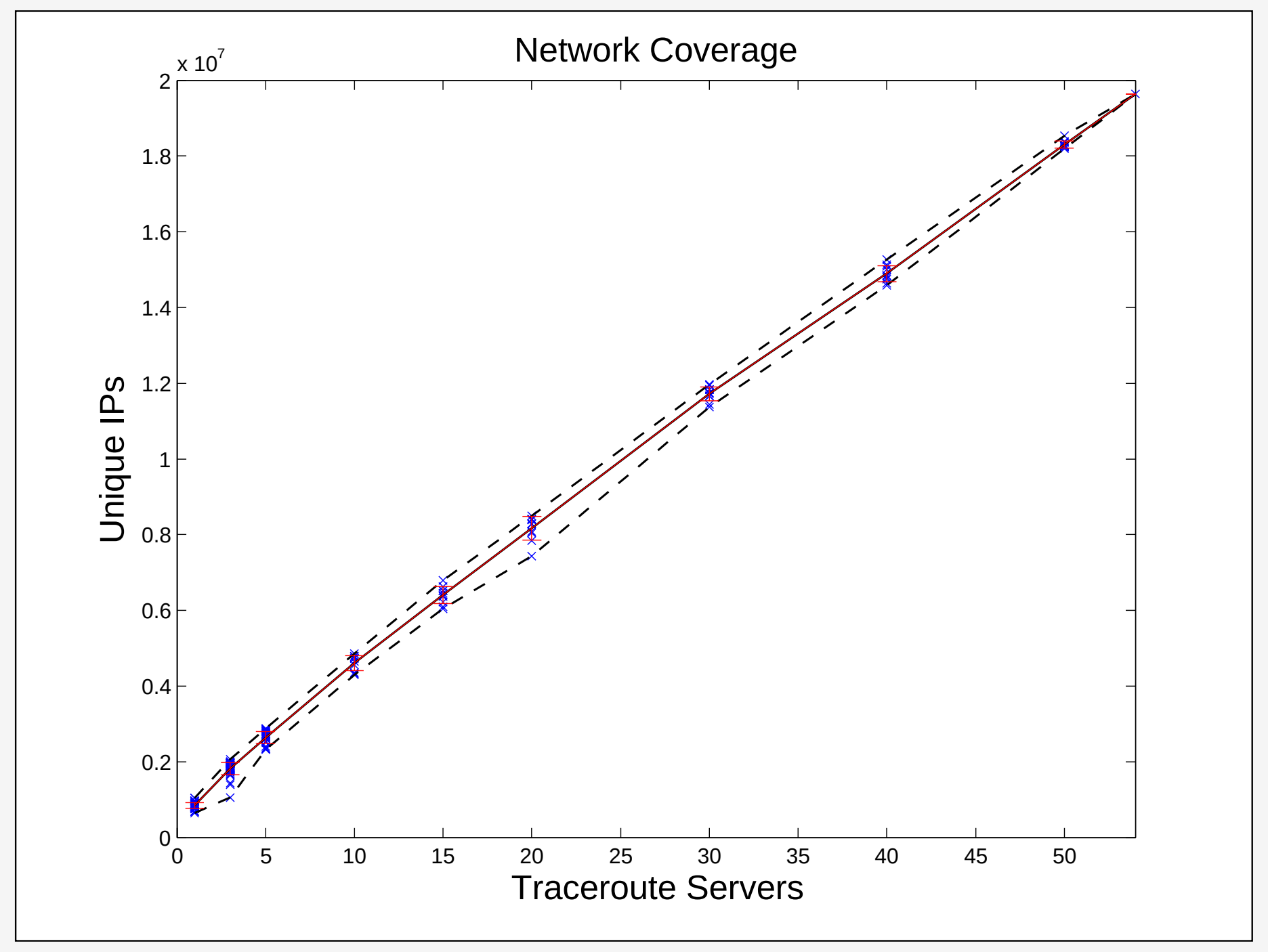
• real-time analytics over continuous streams of data



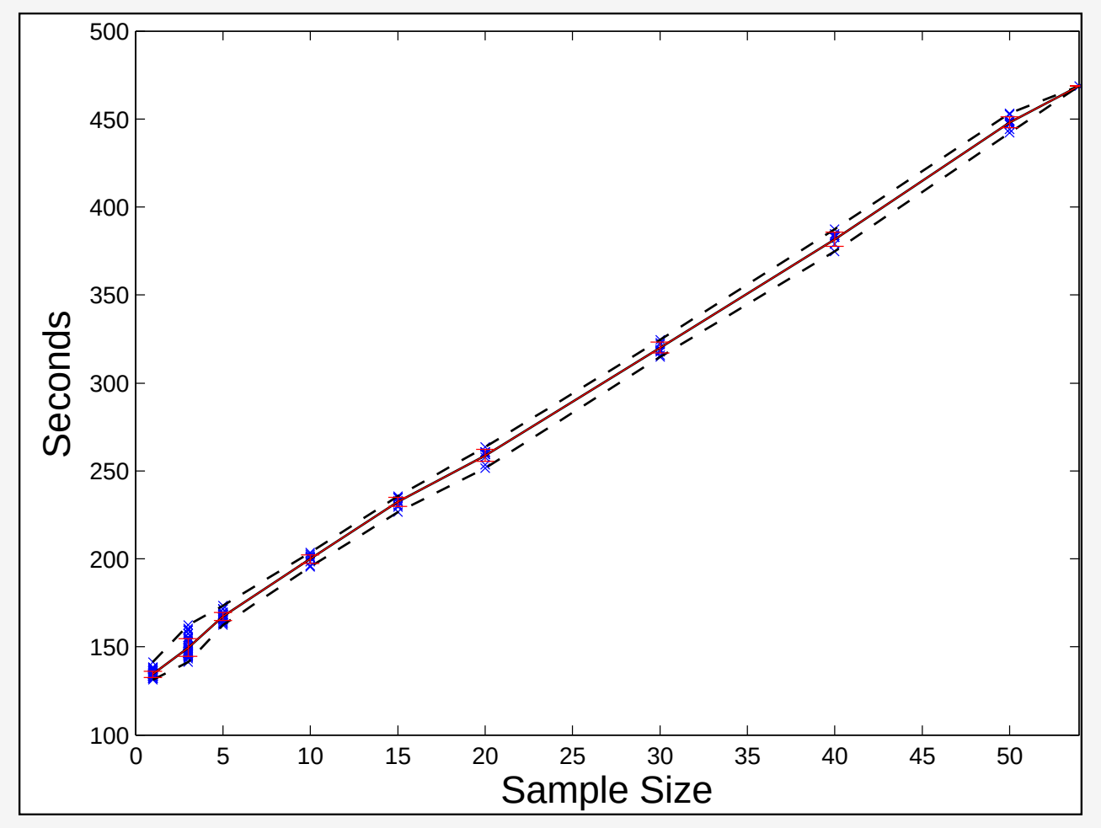
### Hadoop MapReduce – Analysis Example:



## Preliminary Results



**Observations:**
• Traceroute servers see *different* parts of the Internet
• Coverage increases as a function of the *number of servers*
• For coverage to *converge*, one must increase the # of:
  ✓ *Traceroute servers* (locations/data collection points)
  ✓ *Data* (# of traceroute queries from each server)
• Location matters (thus far)



## Future Work

- Increase the number of traceroutes from each server
- Estimate number of traceroutes required for accurate coverage or convergence (from each location(s))
- Model coverage dynamics in real-time
- Analyze coverage using intersection of destination IPs

**References**

1. The IPv4 Routed /24 AS Links Dataset – November 9 – 24, 2011, Young Hyun, Bradley Huffaker, Dan Andersen, Emile Aben, Matthew Luckie, kc claffy, and Colleen Shannon, http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml.

U.S. DEPARTMENT OF ENERGY

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344.

LLNL-POST-571273

Lawrence Livermore National Laboratory